

The Phylogenetic Handbook

A Practical Approach to DNA and Protein Phylogeny

Edited by

Marco Salemi

University of California, Irvine
(Formerly of Katholieke Universiteit Leuven)

and

Anne-Mieke Vandamme

Rega Institute for Medical Research, Katholieke Universiteit Leuven, Belgium



PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS

The Edinburgh Building, Cambridge CB2 2RU, UK

40 West 20th Street, New York, NY 10011-4211, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

Ruiz de Alarcón 13, 28014 Madrid, Spain

Dock House, The Waterfront, Cape Town 8001, South Africa

<http://www.cambridge.org>

© Cambridge University Press 2003

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typefaces Minion 10.5/14 pt. and Formata *System* L^AT_EX 2_ε [TB]

A catalog record for this book is available from the British Library.

Library of Congress Cataloging in Publication Data

The phylogenetic handbook : a practical approach to DNA and protein phylogeny /
edited by Marco Salemi, Anne-Mieke Vandamme.

p. cm.

Includes bibliographical references and index.

ISBN 0-521-80390-X

1. DNA – Analysis – Handbooks, manuals, etc. 2. Proteins – Analysis – Handbooks,
manuals, etc. 3. Cladistic analysis – Handbooks, manuals, etc. I. Salemi, Marco, 1968–II.
Vandamme, Anne-Mieke, 1960–

QP624 .P485 2003

572.8'633 – dc21

2002073927

ISBN 0 521 80390 X hardback

Contents

<i>Foreword</i>	<i>page</i> xvii
<i>Acknowledgments</i>	xxi
<i>Contributors</i>	xxiii
1 Basic concepts of molecular evolution	1
<hr/>	
Anne-Mieke Vandamme	
1.1 Genetic information	1
1.2 Population dynamics	6
1.3 Data used for molecular phylogenetic analysis	10
1.4 What is a phylogenetic tree?	14
1.5 Methods to infer phylogenetic trees	17
1.6 Is evolution always tree-like?	21
2 Sequence databases	24
<hr/>	
THEORY	24
Guy Bottu and Marc Van Ranst	
2.1 General nucleic acid sequence databases	24
2.2 General protein sequence databases	26
2.3 Nonredundant sequence databases	27
2.4 Specialized sequence databases	28
2.5 Databases with aligned protein sequences	29
2.6 Database documentation search	30
2.6.1 Text-string searching	30
2.6.2 Searching by index	30
2.7 ENTREZ database	32
2.8 Sequence similarity searching: BLAST	33
PRACTICE	37
Marco Salemi	
2.9 File formats	37

2.10	Three example data sets	40
2.10.1	Preparing input files: HIV/SIV example data set	41
3	Multiple alignment	45
	THEORY	45
	Des Higgins	
3.1	Introduction	45
3.2	The problem of repeats	46
3.3	The problem of substitutions	47
3.4	The problem of gaps	50
3.5	Testing multiple-alignment methods	51
3.6	Multiple-alignment algorithms	52
3.6.1	Dot-matrix sequence comparison	52
3.6.2	Dynamic programming	54
3.6.3	Genetic algorithms	55
3.6.4	Other algorithms	55
3.7	Progressive alignment	55
3.7.1	Clustal	57
3.7.2	T-Coffee	58
3.8	Hidden Markov models	58
3.9	Nucleotide sequences versus amino-acid sequences	59
	PRACTICE	61
	Des Higgins and Marco Salemi	
3.10	Searching for homologous sequences with BioEdit	61
3.11	File formats for Clustal	63
3.12	Access to ClustalW and ClustalX	64
3.13	Aligning the HIV/SIV sequences with ClustalX	64
3.14	Aligning nucleotide sequences in a coding region with DAMBE	66
3.15	Adding sequences to preexisting alignments	67
3.16	Editing and viewing multiple alignments	68
3.17	Databases of alignments	69
4	Nucleotide substitution models	72
	THEORY	72
	Korbinian Strimmer and Arndt von Haeseler	
4.1	Introduction	72
4.2	Observed and expected distances	73
4.3	Number of mutations in a given time interval *(optional)	74
4.4	Nucleotide substitutions as a homogeneous Markov process	77
4.4.1	The Jukes and Cantor (JC69) model	79
4.5	Derivation of Markov process *(optional)	80
4.5.1	Inferring the expected distances	83

4.6 Nucleotide substitution models	83
4.6.1 Rate heterogeneity over sites	85
PRACTICE: The PHYLIP and TREE-PUZZLE software packages	88
Marco Salemi	
4.7 Software packages	88
4.8 Jukes and Cantor (JC69) genetic distances	90
4.9 Kimura 2-parameters (K80) and F84 genetic distances	91
4.10 More complex models	92
4.10.1 Modeling rate heterogeneity over sites	93
4.11 The problem of substitution saturation	95
4.12 Choosing among different evolutionary models	97
5 Phylogeny inference based on distance methods	101
<hr/>	
THEORY	101
Yves Van de Peer	
5.1 Introduction	101
5.2 Tree-inferring methods based on genetic distances	103
5.2.1 Cluster analysis (UPGMA and WPGMA)	103
5.2.2 Minimum evolution and neighbor-joining	107
5.2.3 Other distance methods	113
5.3 Evaluating the reliability of inferred trees	115
5.3.1 Bootstrap analysis	115
5.3.2 Jackknifing	118
5.4 Conclusions	118
PRACTICE	120
Marco Salemi	
5.5 The TreeView program	120
5.6 Procedure to estimate distance-based phylogenetic trees with PHYLIP	120
5.7 Inferring an NJ tree for the mtDNA data set	121
5.8 Inferring a Fitch-Margoliash tree for the mtDNA data set	125
5.9 Inferring an NJ tree for the HIV-1 data set	125
5.10 Bootstrap analysis with PHYLIP	126
5.11 Other programs	133
6 Phylogeny inference based on maximum-likelihood methods with TREE-PUZZLE	137
<hr/>	
THEORY	137
Arndt von Haeseler and Korbinian Strimmer	
6.1 Introduction	137

6.2	The formal framework	140
6.2.1	The simple case: Maximum-likelihood tree for two sequences	140
6.2.2	The complex case	141
6.3	Computing the probability of an alignment for a fixed tree	142
6.3.1	Felsenstein's pruning algorithm	144
6.4	Finding a maximum-likelihood tree	145
6.4.1	The quartet-puzzling algorithm	146
6.5	Estimating the model parameters with maximum likelihood	149
6.6	Likelihood-mapping analysis	150
	PRACTICE	153
	Arndt von Haeseler and Korbinian Strimmer	
6.7	Software packages	153
6.8	An illustrative example of quartet-puzzling tree reconstruction	153
6.9	Likelihood-mapping analysis of the HIV data set	156
7	Phylogeny inference based on parsimony and other methods using PAUP*	160
	THEORY	160
	David L. Swofford and Jack Sullivan	
7.1	Introduction	160
7.2	Parsimony analysis – background	161
7.3	Parsimony analysis – methodology	163
7.3.1	Calculating the length of a given tree under the parsimony criterion	163
7.4	Searching for optimal trees	166
7.4.1	Exact methods	171
7.4.2	Approximate methods	175
	PRACTICE	182
	David L. Swofford and Jack Sullivan	
7.5	Analyzing data with PAUP* through the command-line interface	182
7.6	Basic parsimony analysis and tree-searching	186
7.7	Analysis using distance methods	193
7.8	Analysis using maximum-likelihood methods	196
8	Phylogenetic analysis using protein sequences	207
	THEORY	207
	Fred R. Opperdoes	
8.1	Introduction	207
8.2	Why protein sequences?	209
8.2.1	The genetic code	210

8.2.2 Codon bias	210
8.2.3 Long time horizon	210
8.2.4 Phylogenetic noise reduction	211
8.2.5 Introns and noncoding DNA	211
8.2.6 Multigene families and post-transcriptional editing	212
8.3 Measurement of sequence divergence in proteins: The PAM	213
8.4 Alignment of protein sequences	215
8.4.1 Sequence retrieval and multiple-sequence alignment	219
8.4.2 Secondary-structure-based alignment	219
8.4.3 Prodom, Pfam, and Blocks databases	220
8.4.4 Manual adjustment of a protein alignment	220
8.5 Tree-building methods for protein phylogeny	221
8.6 Some good advice	224
PRACTICE	226
Fred R. Opperdoes	
8.7 A phylogenetic analysis of the Leishmanial GPD gene carried out via the Internet	226
8.8 A comparison of the trypanosomatid phylogeny from nucleotide and protein sequences	230
8.9 Implementing different evolutionary models with DAMBE and TREE-PUZZLE	233
9 Analysis of nucleotide sequences using TREECON	236
THEORY	236
Yves Van de Peer	
9.1 Introduction	236
9.2 TREECON, distance trees, and among-site rate variation	236
9.2.1 Taking into account among-site rate variation: An example	241
9.3 Conclusions	245
PRACTICE	246
Yves Van de Peer	
9.4 The TREECON software package	246
9.5 Implementation	246
9.6 Substitution rate calibration	251
10 Selecting models of evolution	256
THEORY	256
David Posada	
10.1 Models of evolution and phylogeny reconstruction	256
10.2 The relevance of models of evolution	257

10.3	Selecting models of evolution	257
10.4	The likelihood ratio test	258
10.4.1	LRTs and parametric bootstrapping	259
10.4.2	Hierarchical LRTs	260
10.4.3	Dynamical LRTs	261
10.5	Information criteria	263
10.5.1	AIC	264
10.5.2	BIC	264
10.6	Fit of a single model to the data	264
10.7	Testing the molecular clock hypothesis	265
10.7.1	The relative rate test	266
10.7.2	LRT of the global molecular clock	267
	PRACTICE	270
	David Posada	
10.8	The model-selection procedure	270
10.9	The program MODELTEST	273
10.10	Implementing the LRT of the molecular clock using PAUP*	275
10.11	Selecting the best-fit model in the example data sets	276
10.11.1	Vertebrate mtDNA	277
10.11.2	HIV envelope gene	278
10.11.3	G3PDH protein	279
11	Analysis of coding sequences	283
	THEORY	283
	Yoshiyuki Suzuki and Takashi Gojobori	
11.1	Introduction	283
11.2	Mutation fraction methods	285
11.2.1	Method of Nei and Gojobori (NG86 method)	285
11.2.2	Method of Zhang et al. (ZRN98 method)	287
11.2.3	Method of Ina (I95 method)	288
11.3	Degenerate site methods	290
11.3.1	Method of Li et al. (LWL85 method)	291
11.3.2	Method of Pamilo and Bianchi, and Li (PBL93 method)	294
11.4	Codon model methods	294
11.4.1	Method of Muse (M96 method)	295
11.4.2	Method of Yang and Nielsen (YN98 method)	296
11.5	Methods for estimating d_S and d_N at single codon sites	296
11.5.1	Method of Suzuki and Gojobori (SG99 method)	297
11.6	Test of neutrality for two sequences	298
11.6.1	Z test	298
11.6.2	Likelihood ratio test (LRT)	298
11.6.3	Window analysis	299

11.7 Test of neutrality at single codon sites	299
11.7.1 Method of Nielsen and Yang (1998) (NY98 method)	300
11.7.2 SG99 method	300
PRACTICE	302
Yoshiyuki Suzuki and Takashi Gojobori	
11.8 Software for analyzing coding sequences	302
11.9 Estimation of d_S and d_N in an HCV data set	302
11.9.1 Estimation of d_S and d_N with NG86, ZRN98, LWL85, and PBL93 methods (MEGA2)	303
11.9.2 Estimation of d_S and d_N with YN98 method (PAML)	304
11.9.3 Comparing different estimates of d_S and d_N	305
11.10 An example of window analysis	306
11.11 Detection of positive selection at single amino acid sites	307
11.12 Conclusions	308
12 SplitsTree: A network-based tool for exploring evolutionary relationships in molecular data	312
<hr/>	
THEORY	312
Vincent Moulton	
12.1 Exploring evolutionary relationships through networks	312
12.2 An introduction to split-decomposition theory	314
12.2.1 The Buneman tree	315
12.2.2 Split decomposition	316
12.3 From weakly compatible splits to networks	318
PRACTICE	320
Vincent Moulton	
12.4 The SplitsTree program	320
12.5 Using SplitsTree on the mtDNA data set	320
12.6 Using SplitsTree on the HIV-1 data set	324
13 Tetrapod phylogeny and data exploration using DAMBE	329
<hr/>	
THEORY	329
Xuhua Xia and Zheng Xie	
13.1 The phylogenetic problem and the sequence data	329
13.2 Results of routine phylogenetic analyses without data exploration	330
13.3 Distance-based statistical test of alternative phylogenetic trees (optional)	332
13.4 Likelihood-based statistical tests of alternative phylogenetic trees	333

13.5 Data exploration	335
13.5.1 Nucleotide frequencies	335
13.5.2 Substitution saturation and the rate heterogeneity over sites	337
13.5.3 The pattern of nucleotide substitution	338
13.5.4 Insertion and deletion as phylogenetic characters	339
PRACTICE	342
Xuhua Xia and Zheng Xie	
13.6 Data exploration with DAMBE	342
13.6.1 Nucleotide frequencies	342
13.6.2 Basic phylogenetic reconstruction	342
13.6.3 Rate heterogeneity over sites estimated through reconstruction of ancestral sequences	343
13.6.4 Empirical substitution pattern	344
13.6.5 Testing alternative phylogenetic hypotheses with the distance-based method	344
13.6.6 Testing alternative phylogenetic hypotheses with the likelihood-based method	345
14 Detecting recombination in viral sequences	348
THEORY	348
Mika Salminen	
14.1 Introduction and theoretical background to exploring recombination in viral sequences	348
14.2 Requirements for detecting recombination	349
14.3 Theoretical basis for methods to detect recombination	351
14.4 Examples of viral recombination	360
PRACTICE	362
Mika Salminen	
14.5 Existing tools for analysis of recombination	362
14.6 Analyzing example sequences to visualize recombination	364
14.6.1 Exercise 1: Working with <code>Simplot</code>	364
14.6.2 Exercise 2: Mapping recombination with <code>Simplot</code>	368
14.6.3 Exercise 3: Using the “groups” feature of <code>Simplot</code>	369
14.6.4 Exercise 4: Using <code>SplitsTree</code> to visualize recombination	373
15 LAMARC: Estimating population genetic parameters from molecular data	378
THEORY	378
Mary K. Kuhner	
15.1 Introduction	378

15.2 Basis of the Metropolis-Hastings MCMC sampler	379
15.2.1 Random sample	381
15.2.2 Stability	381
15.2.3 No other forces	381
15.2.4 Evolutionary model	381
15.2.5 Large population relative to sample	382
15.2.6 Adequate run time	382
PRACTICE	384
Mary K. Kuhner	
15.3 The LAMARC software package	384
15.3.1 FLUCTUATE (COALESCE)	384
15.3.2 MIGRATE	384
15.3.3 RECOMBINE	385
15.3.4 LAMARC	386
15.4 Starting values	386
15.5 Space and time	387
15.6 Sample size considerations	387
15.7 Virus-specific issues	388
15.7.1 Multiple loci	388
15.7.2 Rapid growth rates	388
15.7.3 Sequential samples	389
15.8 An exercise with LAMARC	389
15.8.1 Exercise using FLUCTUATE	390
15.8.2 Exercise using RECOMBINE	395
15.9 Conclusions	396
<i>Index</i>	399
<i>Color section follows p. 328.</i>	

Foreword

Theodosius Dobzhansky (1973) wisely said, “Nothing in biology makes sense except in the light of Evolution.” This truism is so often repeated that it is nearly a mantra and, with the complete genomes of many organisms being completed nearly daily, all kinds of people, but especially molecular biologists and informaticists, are rediscovering that truth. And with that discovery they are coming to need to know the tools of the trade that have been under development for nearly forty years. This book is for them in particular but it has much that, except for polymaths, may be useful even to the cognoscenti.

The book has grown out of Drs. Vandamme’s and Salemi’s annual course in these methods at the Katholieke Universiteit in Leuven, Belgium, where they have produced an exceptional workshop for eight years that does for Europe what a similar outstanding and long-running workshop at Woods Hole did for the United States and Canada. But the latter has not created a book like this.

The coverage is comprehensive. Topics touched upon include databases, multiple alignments, nucleotide substitution models, phylogeny inference methods (such as distance, maximum likelihood, and maximum parsimony), post-phylogenetic information (such as molecular clocks and selection), and useful subsidiary statistical techniques (such as bootstrapping and likelihood ratio tests).

Each of the major sections is written by an expert in the field, and each such section is divided into two major subsections, theory and practice. This permits the novice to proceed with his analysis without having to master the theory. That is, of course, very dangerous in this field where so many methods have different assumptions and the failure of any one of those assumptions (clocklike behavior, all sites equally mutable, all substitutions neutral) can reduce your analysis to rubbish, if untrue, which they frequently are. Still, there are people like that and we may hope that a good text such as this, with its many caveats and generally simple prose, will reduce the published trash.

The material is enhanced by the use of specific examples from which you can see what to expect, and see if you can get the same answer, and then try your own data

to see if anything strange has happened. The examples also aid in locating what you need to find in the text.

Another aspect of the book that enhances its utility for the reader is the repeated use of the same three data sets, even by different authors, to illustrate the methods. This increases immensely the value of the exercises. This is especially true when the results from different methods are ostensibly for the same desired end, and one gets to see how they differ and why (or at least to worry about it).

The example data sets used in the book can be downloaded from the book's website [<http://www.kuleuven.ac.be/aidslab/phylogenybook.htm>]. On the website the reader can also find useful links to the major phylogeny resources on the internet, as well as the results of all the analyses discussed in the text, including phylogenetic trees, unaligned and aligned sequences, and so forth.

It is appropriate to compare this work with others in the general area. The first two are by Weir (1990) and by Waterman (1995). They are both highly theoretical and quite capable of turning off many biologists quickly (although Waterman's book can be highly engaging as in his recounting of the efforts of George Gamow to predict that the genetic code was a commaless code). At the other extreme, Hall (2001) is really simple-minded enough (intentionally so) that a bright senior could easily master the methods. However, the Hall book lacks the comprehensiveness of the Salemi and Vandamme work. Two other good books, Li (1998), and Page and Holmes (1998), are largely theoretical although they make a great effort to make the subject palatable to the biologist who is mathematically challenged. In sum, there is no other book even trying to occupy the niche of this one.

In conclusion, this is a relatively easy-to-use workbook for phylogenetics, especially if the index is properly looked to (I haven't seen it). However, I have to present a strongly worded negative comment. Although tables and figures in the book have titles, many have no legends and many of the remainder have poor legends. For example, numbers normally have dimensions, (such as nucleotide differences per hundred nucleotide positions), that should have been given. Figures and tables should be as self-sufficient as is reasonable. This is not true here. Let us hope this is corrected in the next printing, which I am sure this book will achieve.

Walter M. Fitch
December 27, 2002

REFERENCES

- Dobzhansky, Theodosius (1973). Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*, 125–129.
- Weir, Bruce S. (1990). *Genetic Data Analysis*, Sunderland, MA: Sinauer Associates.

- Waterman, Michael S. (1995). *Introduction to Computational Biology: Maps, sequences and genomes*.
New York: Chapman and Hall.
- Hall, Barry G. (2001). *Phylogenetic Trees Made Easy*, Sunderland, MA: Sinauer Associates.
- Li, Wen-Hsiung (1997). *Molecular Evolution*, Sunderland, MA: Sinauer Associates.
- Page, Roderic D. M. and Edward C., Holmes (1998). *Molecular Evolution: A Phylogenetic Approach*.
London: Blackwell Science.

Note: During the writing of this book the alpha release of the new version of PHYLIP, PHYLIP 3.6 has been made available on the PHYLIP web page. All the exercises with PHYLIP refer to version 3.5, but additional exercises covering PHYLIP v3.6 can be found on the book website.

Contributors

Guy Bottu

Belgian EMBnet Node
Brussels, Belgium

Walter Fitch (Foreword)

Ecology and Evolutionary Biology
University of California, Irvine
California, USA

Takashi Gojobori

Center for Information Biology
and DNA Data Bank of Japan
National Institute of Genetics
Mishima, Japan

Arndt von Haeseler

Heinrich-Heine-Universität Düsseldorf
Institut für Bioinformatik
Düsseldorf, Germany

Des Higgins

Department of Biochemistry
University College
Cork, Ireland

Mary K. Kuhner

Department of Genome Sciences
University of Washington
Washington, USA

Vincent Moulton

Physics and Mathematics Department
Mid Sweden University
Sundsvall, Sweden

Fred R. Opperdoes

C. de Duve Institute of Cellular Pathology
Université Catholique de Louvain
Brussels, Belgium

Yves Van de Peer

Department of Plant Systems Biology
Flanders Interuniversity Institute for
Biotechnology (VIB)
Ghent University
Ghent, Belgium

David Posada

Departamento de Bioquímica, Xenética e
Immunoloxía
Facultade de Ciencias
Universidade de Vigo
Vigo, Spain

Marc Van Ranst

Rega Institute for Medical Research
Katholieke Universiteit Leuven
Leuven, Belgium

Marco Salemi

Rega Institute for Medical Research
Katholieke Universiteit Leuven
Leuven, Belgium
and
Ecology and Evolutionary Biology
University of California, Irvine
California, USA

Mika Salminen

HIV Laboratory
National Public Health Institute
Department of Infectious Disease
Epidemiology
Helsinki, Finland

Korbinian Strimmer

Department of Statistics
University of München
München, Germany

Jack Sullivan

Department of Biological Science
University of Idaho
Idaho, USA

Yoshiyuki Suzuki

Center for Information Biology
and DNA Data Bank of Japan
National Institute of Genetics
Mishima, Japan

David L. Swofford

School of Computational Science and
Information Technology
and Department of Biological Science
Florida State University
Florida, USA

Anne-Mieke Vandamme

Rega Institute for Medical Research
Katholieke Universiteit Leuven
Leuven, Belgium

Xuhua Xia

Biology Department
University of Ottawa
Ottawa, Ontario
Canada

Zheng Xie

Institute of Environmental Protection
Hunan University
China

Phylogeny inference based on distance methods

THEORY

Yves Van de Peer

5.1 Introduction

In addition to *maximum parsimony* (*MP*) and likelihood methods (see Chapters 6 and 7), pairwise *distance methods* form the third large group of methods to infer evolutionary trees from sequence data (Figure 5.1). In principle, distance methods try to fit a tree to a matrix of pairwise *genetic distances* (Felsenstein, 1988). For every two sequences, the distance is a single value based on the fraction of positions in which the two sequences differ, defined as *p-distance* (see Chapter 4). The *p-distance* is an underestimation of the true genetic distance because some of the aligned nucleotides are the result of multiple events. Indeed, because mutations are fixed in the genes, there has been an increasing chance of multiple substitutions occurring during evolution at the same sequence position. Therefore, in distance-based methods, one tries to estimate the number of substitutions that have actually occurred by applying a specific *evolutionary model* that makes assumptions about the nature of evolutionary changes (see Chapter 4). When all the pairwise distances have been computed for a set of sequences, a tree topology can then be inferred by a variety of methods (Figure 5.2).

Correct estimation of the genetic distance is crucial and, in most cases, more important than the choice of method to infer the tree topology. Using an unrealistic evolutionary model can cause serious artifacts in tree topology, as previously shown in numerous studies (e.g., Olsen, 1987; Lockhart et al., 1994; Van de Peer et al., 1996; see also Chapter 10). However, because the exact historical record of events that occurred in the evolution of sequences is not known, the best method for estimating the genetic distance is not necessarily self-evident (see Chapter 9).

	Character-based methods	Noncharacter-based methods
Methods based on an explicit model of evolution	Maximum-likelihood methods	Pairwise-distance methods
Methods not based on an explicit model of evolution	Maximum-parsimony methods	

Figure 5.1 Pairwise distance methods are non-character-based methods that make use of an explicit substitution model.

Substitution models are discussed in Chapters 4, 9, and 10. Chapters 7 and 10 discuss how to select the best-fitting evolutionary model for a given data set of nucleotide or amino-acid aligned sequences in order to get an accurate estimation of the genetic distances. In the following sections, it is assumed that genetic distances were estimated using an appropriate evolutionary model, and some of the methods used for inferring tree topologies on the basis of these distances are briefly outlined. However, by no means should this be considered a complete discussion of distance methods; additional discussions are in Felsenstein (1982), Swofford et al. (1996), Li (1997), and Page and Holmes (1998).

Step 1
Estimation of evolutionary distances

```

C T T C A A T C A G G C C C G A
  | | | | | | | | | | | |
A T C A A G T C A G G T T C G A
  | | | | | | | | | | | |
B T C C A G T T A G A C T C G A
  | | | | | | | | | | | |
C T T C A A T C A G G C C C G A
    
```

Convert dissimilarity into evolutionary distance by correcting for multiple events per site (e.g., Jukes and Cantor, 1969):

$$d_{AB} = -\frac{3}{4} \ln \left(1 - \frac{4}{3} \cdot 0.266 \right) = 0.328$$

	A	B	C
B	0.267		
C	0.333	0.333	

dissimilarities

↓

	A	B	C
B	0.330		
C	0.441	0.441	

evolutionary distances

Step 2
Infer tree topology on the basis of estimated evolutionary distances

Figure 5.2 Distance methods proceed in two steps. First, the evolutionary distance is computed for every sequence pair. Usually, this information is stored in a matrix of pairwise distances. Second, a tree topology is inferred on the basis of the specific relationships between the distance values.

5.2 Tree-inferring methods based on genetic distances

The main distance-based tree-building methods are cluster analysis and minimum evolution. Both rely on a different set of assumptions, and their success or failure in retrieving the correct phylogenetic tree depends on how well any particular data set meets such assumptions.

5.2.1 Cluster analysis (UPGMA and WPGMA)

Clustering methods are tree-building methods that were originally developed to construct taxonomic phenograms (Sokal and Michener, 1958; Sneath and Sokal, 1973); that is, trees based on overall phenotypic similarity. Later, these methods were applied to phylogenetics to construct *ultrametric trees*. *Ultrametricity* is satisfied when, for any three *taxa*, A, B, and C,

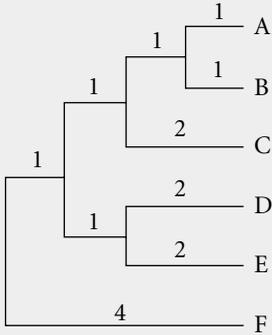
$$d_{AC} \leq \max(d_{AB}, d_{BC}). \quad (5.1)$$

In practice, Equation 5.1 is satisfied when two of the three distances under consideration are equal and as large (or larger) as the third one. Ultrametric trees are *rooted trees* in which all the end nodes are equidistant from the root of the tree, which is only possible by assuming a *molecular clock* (see Chapters 1 and 10). Clustering methods such as the *unweighted-pair group method with arithmetic means* (UPGMA) or the *weighted-pair group method with arithmetic means* (WPGMA) use a sequential clustering algorithm. A tree is built in a stepwise manner, by grouping sequences or groups of sequences – usually referred to as *operational taxonomic units* (OTUs) – that are most similar to each other; that is, for which the genetic distance is the smallest. When two OTUs are grouped, they are treated as a new single OTU (Box 5.1). From the new group of OTUs, the pair for which the similarity is highest is again identified, and so on, until only two OTUs are left. The method applied in Box 5.1 is actually the WPGMA, in which the averaging of the distances is not based on the total number of OTUs in the respective clusters. For example, when OTUs A, B (which have been grouped before), and C are grouped into a new node ‘*u*’, then the distance from node ‘*u*’ to any other node ‘*k*’ (e.g., grouping D and E) is computed as follows:

$$d_{uk} = \frac{d_{(A,B)k} + d_{Ck}}{2} \quad (5.2)$$

Conversely, in UPGMA, the averaging of the distances is based on the number of OTUs in the different clusters; therefore, the distance between ‘*u*’ and ‘*k*’ is

Box 5.1 Cluster analysis (Sneath and Sokal, 1973)



	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

N = 6

Cluster analysis proceeds as follows:

1. Group together (cluster) these OTUs for which the distance is minimal; e.g., A and B together. The depth of the divergence is the distance between A and B divided by 2.



2. Compute the distance from cluster (A, B) to every other OTU.

$$d_{(AB)C} = (d_{AC} + d_{BC})/2 = 4$$

$$d_{(AB)D} = (d_{AD} + d_{BD})/2 = 6$$

$$d_{(AB)E} = (d_{AE} + d_{BE})/2 = 6$$

$$d_{(AB)F} = (d_{AF} + d_{BF})/2 = 8$$

	(AB)	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

Repeat Steps 1 and 2 until all OTUs are clustered (repeat until N = 2).

$$N = N - 1 = 5$$

1. Group together (cluster) these OTUs for which the distance is minimal; e.g., group D and E together. Alternatively, (AB) could be grouped with C.

